Unsupervised Sentiment Analysis for Social Media Images

Yilin Wang, Suhang Wang, Jiliang Tang, Huan Liu, and Baoxin Li
Arizona State University
Tempe, Arizona
{yilin.wang.1, suhang.wang, jiliang.tang, huan.liu, baoxin.li}@asu.edu

Abstract

Recently text-based sentiment prediction has been extensively studied, while image-centric sentiment analysis receives much less attention. In this paper, we study the problem of understanding human sentiments from large-scale social media images, considering both visual content and contextual information, such as comments on the images, captions, etc. The challenge of this problem lies in the "semantic gap" between low-level visual features and higher-level image sentiments. Moreover, the lack of proper annotations/labels in the majority of social media images presents another challenge. To address these two challenges, we propose a novel Unsupervised SEntiment Analysis (USEA) framework for social media images. Our approach exploits relations among visual content and relevant contextual information to bridge the "semantic gap" in the prediction of image sentiments. With experiments on two large-scale datasets, we show that the proposed method is effective in addressing the two challenges.

1 Introduction

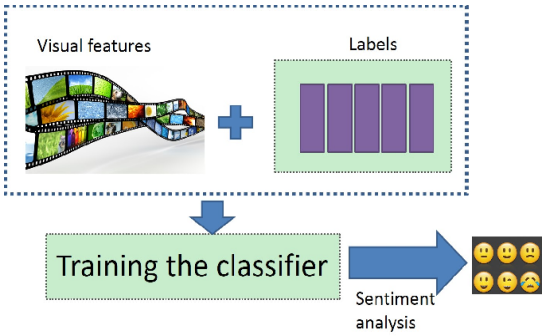
Recent years have witnessed the explosive popularity of image-sharing services such as Flickr[1] and Instagram[2]. For example, as of 2013, 87 millions of users have registered with Flickr[3]. Also, it was estimated that about 20 billion Instagram photos are shared to 2014[4]. Since by sharing photos, users could also express opinions or sentiments, social media images provide a potentially rich source for understanding public opinions/sentiments. Such an understanding may in turn benefit or even enable many real-world applications such as advertisement, recommendation, marketing and health-care. The importance of sentiment analysis for social media images has thus attracted increasing attention recently [Yang *et al.*, 2014; You *et al.*, 2015].

[1] www.flickr.com
[2] www.instagram.com
[3] http://en.wikipedia.org/wiki/Flickr
[4] http://blog.instagram.com/post/80721172292/200m

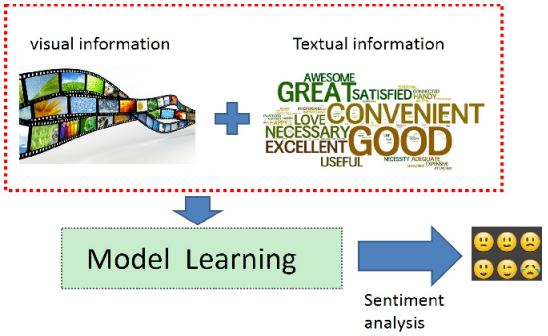
Current methods of sentiment analysis for social media images include low-level visual feature based approaches [Jia *et al.*, 2012; Yang *et al.*, 2014], mid-level visual feature based approaches [Borth *et al.*, 2013; Yuan *et al.*, 2013] and deep learning based approaches [You *et al.*, 2015]. The vast majority of existing methods are supervised, relying on labeled images to train sentiment classifiers. Unfortunately, sentiment labels are in general unavailable for social media images, and it is too labor- and time-intensive to obtain labeled sets large enough for robust training. In order to utilize the vast amount of unlabeled social media images, an unsupervised approach would be much more desirable. This paper studies *unsupervised sentiment analysis*.

Typically, visual features such as color histogram, brightness, the presence of objects and visual attributes lack the level of semantic meanings required by sentiment prediction. In supervised case, label information could be directly utilized to build the connection between the visual features and the sentiment labels. Thus, unsupervised sentiment analysis for social media images is inherently more challenging than its supervised counterpart. As images from social media sources are often accompanied by textual information, intuitively such information may be employed. However, textual information accompanying images is often incomplete (e.g., scarce tags) and noisy (e.g., irrelevant comments), and thus often inadequate to support independent sentiment analysis [Hu and Liu, 2004; Hu *et al.*, 2013b]. On the other hand, such information can provide much-needed additional semantic information about the underlying images, which may be exploited to enable unsupervised sentiment analysis. How to achieve this is the objective of our approach.

In this paper, we study unsupervised sentiment analysis for social media images with textual information by investigating two related challenges: (1) how to model the interaction between images and textual information systematically so as to support sentiment prediction using both sources of information, and (2) how to use textual information to enable unsupervised sentiment analysis for social media images. In addressing these two challenges, we propose a novel Unsupervised SEntiment Analysis (USEA) framework, which performs sentiment analysis for social media images in an unsupervised fashion. Figure 1 schematically illustrates the difference between the proposed unsupervised method and existing supervised methods. Supervised methods use label informa-



(a) Supervised Sentiment Analysis.



(b) The Proposed Unsupervised Sentiment Analysis.

Figure 1: Sentiment Analysis for Social Media Images.

tion to learn a sentiment classifier; while the proposed method does not assume the availability of label information but employ auxiliary textual information. Our main contribution can be summarized as below:

- A principled approach to enable unsupervised sentiment analysis for social media images.
- A novel unsupervised sentiment analysis framework USEA for social media images, which captures visual and textual information into a unifying model. To our best knowledge, USEA is the first unsupervised sentiment analysis framework for social media images; and
- Comparative studies and evaluations using datasets from real-world social media image-sharing sites, documenting the performance of USEA and leading existing methods, serving as benchmark for further exploration.

2 Problem Statement

In this paper, scalars are denoted by lower-case letters (a, b, . . . ; α, β, . . .), vectors are written as lower-case bolded letters ($\mathbf{a}$, $\mathbf{b}$, . . .), and matrices correspond to boldfaced uppercase letters ($\mathbf{A}$, $\mathbf{B}$, . . .). Let $\mathcal{I} = \{I_1, I_2, \ldots, I_n\}$ be the set of images where n is the number of images. We use $\mathcal{P} = \{p_1, p_2, \ldots, p_n\}$ to denote associated textual information about images where p_i is the textual information about I_i. Let $\mathcal{F}_v$ be set of m_v visual features and $\mathcal{F}_t$ be set of m_t textual features. We use $\mathbf{X}_v \in \mathbb{R}^{n\times m_v}$ and $\mathbf{X}_t \in \mathbb{R}^{n\times m_t}$ to denote visual and textual information about images, respectively. Let $\mathcal{C} = \{c_1, c_2, \ldots, c_k\}$ be the set of sentiment labels. Note that in this work we only consider positive, neutral and negative sentiments with $k = 3$ but the generalization of the proposed framework to multi-class sentiment analysis is straightforward.

With the aforementioned notations/definitions, the problem of unsupervised sentiment analysis for social media images with textual information is formally defined as:

Given n images with visual information $\mathbf{X}_v$ and textual information $\mathbf{X}_t$, to predict sentiment labels in $\mathcal{C}$ for the given n images.

3 Unsupervised Sentiment Analysis for Social Media Images

In this section, we first present our method for exploiting text information and then introduce the unsupervised sentiment analysis framework with an optimization method.

3.1 Exploiting Textual Information

Without label information, it is challenging for unsupervised sentiment analysis to connect visual features with sentiment labels. Textual information associated with social media images may be exploited to help, as it provides semantics about the underly images and in particular rich sentiment signals such as sentiment words and emotion symbols may be found in the textual fields. Hence, to exploit textual information, we investigate (1) how to incorporate textual information into visual information; and (2) how to model sentiment signals in textual information.

Since visual and textual information are two views about the same set of images, it is reasonable to assume that they share the same sentiment label space. More specifically, the sentiment of I_i should be consistent with that of its associated textual information p_i. Let $\mathbf{U}_0 \in \mathbb{R}^{n\times k}$ be the sentiment label space where $\mathbf{U}_0(i,j) = 1$ if the i-th data instance belongs to c_j, and $\mathbf{U}_0(i,j) = 0$ otherwise. We propose the following formulation to incorporate visual information with textual information based on nonnegative matrix factorization:

$$\begin{aligned} \min_{\mathbf{UV}} \; & \left\|\mathbf{X}_v - \mathbf{U}_v\mathbf{V}_v^T\right\|_F^2 + \alpha\left\|\mathbf{X}_t - \mathbf{U}_t\mathbf{V}_t^T\right\|_F^2 \\ & + \beta(\|\mathbf{U}_v - \mathbf{U}_0\|_F^2 + \|\mathbf{U}_t - \mathbf{U}_0\|_F^2) \\ \text{subject to} \quad & \mathbf{U}_v \geq 0, \mathbf{U}_t \geq 0; ||\mathbf{U}_0(i,:)||_0 = 1, i \in \{1,2,..n\} \\ & \mathbf{U}_0(i,j) \in \{0,1\}\, j \in \{1,2,..k\} \end{aligned} \tag{1}$$

where α controls how textual information contributes to the model and $||\cdot||_0$ is ℓ_0, which counts the number of nonzero entries in the vector. $\mathbf{U}_v \in \mathbb{R}^{n\times k}$ and $\mathbf{U}_t \in \mathbb{R}^{n\times k}$ are the sentiment label spaces learned from visual information and textual information, respectively. The term of $\beta(\|\mathbf{U}_v - \mathbf{U}_0\|_F^2 + \|\mathbf{U}_t - \mathbf{U}_0\|_F^2)$ ensures that these two types of information should share the sentiment label space $\mathbf{U}_0$. $\mathbf{V}_v \in \mathbb{R}^{m_v\times k}$ and $\mathbf{V}_t \in \mathbb{R}^{m_t\times k}$ indicate the sentiment polarities of visual and textual features, respectively.

Textual information contains rich sentiment signals. First, some words may contain sentiment polarities. For example,

some words are positive such as "happy" and "terrific"; while others are negative such as "gloomy" and "disappointed". The sentiment polarities of words can be obtained via some public sentiment lexicons. For example, the sentiment lexicon MPQA contains 7,504 human labeled words which are commonly used in the daily life with 2,721 positive words and 4,783 negative words. Second, some abbreviations and emoticons are strong sentiment indicators. For example, "lol"(means laughing out loud) is a positive indicator while ":(" is a negative indicator. Let $\mathbf{V}_{t0} \in \mathbb{R}^{m_v \times k}$ be the matrix coding sentiment signals in textual information where $\mathbf{V}_{t0}(i,j) = 1$ if i-th word belongs to c_j and $\mathbf{V}_{t0}(i,j) = 0$ otherwise. To model sentiment signals, we force the learned sentiment polarities of textual features to be consistent with those indicated by sentiment signals. Furthermore, not all textual features in $\mathcal{F}_t$ contain sentiment polarities and $\mathbf{V}_t$ should be sparse. We propose the following formulation to achieve these two goals as:

$$\min \quad \|\mathbf{V}_t - \mathbf{V}_{t0}\|_{2,1} \tag{2}$$

$\|\mathbf{X}\|_{2,1}$ is the $\ell_{2,1}$ of the matrix $\mathbf{X}$, which ensures the row sparsity of $\mathbf{X}$ [Nie *et al.*, 2010].

The significance of textual information in unsupervised sentiment analysis for social media images is two-fold. First, textual information bridges the semantic gap between visual features and sentiment labels. Second, we are allowed to do sentiment analysis for social media images in an unsupervised scenarios by modeling textual information via Eqs. (1) and (2).

3.2 The Framework: USEA

By combining the above discussion, we can have the following initial framework, which provides a potential solution to inferring sentiments by jointly considering visual information and corresponding contextual information:

$$\begin{aligned} \min_{\mathbf{UV}} & \left\|\mathbf{X}_v - \mathbf{U}_v\mathbf{V}_v^T\right\|_F^2 + \alpha\left\|\mathbf{X}_t - \mathbf{U}_t\mathbf{V}_t^T\right\|_F^2 \\ & + \beta(\|\mathbf{U}_v - \mathbf{U}_0\|_F^2 + \|\mathbf{U}_t - \mathbf{U}_0\|_F^2) \\ & + \gamma\|\mathbf{V}_t - \mathbf{V}_{t0}\|_{2,1} \\ \text{s.t.} \quad & \mathbf{U}_v \geq 0; \mathbf{U}_t \geq 0, ||\mathbf{U}_0(i,:)||_0 = 1, i \in \{1,2,..n\} \\ & \mathbf{U}_0(i,j) \in \{0,1\}\, j \in \{1,2,..k\} \end{aligned} \tag{3}$$

The parameter γ controls the sparsity of regularization term. However, the constrains of $\mathbf{U}_0$ in Eq. (3), mixed vector zero norm with integer programming, make the problem difficult to solve. To tackle this problem, we consider the relaxation of $\mathbf{U}_0$ by adding the extra orthogonal constraint on the value of $\mathbf{U}_0$. With the relaxation, the proposed framework (USEA) is to solve the following optimization problem:

$$\begin{aligned} \min_{\mathbf{UV}} & \left\|\mathbf{X}_v - \mathbf{U}_v\mathbf{V}_v^T\right\|_F^2 + \alpha\left\|\mathbf{X}_t - \mathbf{U}_t\mathbf{V}_t^T\right\|_F^2 \\ & + \beta(\|\mathbf{U}_v - \mathbf{U}_0\|_F^2 + \|\mathbf{U}_t - \mathbf{U}_0\|_F^2) \\ & + \gamma\|\mathbf{V}_t - \mathbf{V}_{t0}\|_{2,1} \\ \text{s.t.} \quad & \mathbf{U}_v \geq 0, \quad \mathbf{U}_t \geq 0, \\ & \mathbf{U}_0^T\mathbf{U}_0 = \mathbf{I}; \mathbf{U}_0 \geq 0 \end{aligned} \tag{4}$$

3.3 An Optimization Method

There are 5 components, i.e. $\mathbf{U}_v, \mathbf{V}_v, \mathbf{U}_t, \mathbf{V}_t$ and $\mathbf{U}_0$, in Eq. (4). Thus it is difficult to optimize all the components simultaneously. In the following parts, we demonstrate an alternating algorithm to optimize the objective function by updating each component iteratively.

Update $\mathbf{V}_t$: If $\mathbf{U}_0$, $\mathbf{U}_v$, $\mathbf{V}_v$ and $\mathbf{U}_t$ are fixed, then the objective function is decoupled and the constrains are independent of $\mathbf{V}_t$. Thus we can optimize $\mathbf{V}_t$ separately and ignore the term without V_t, leading to the following:

$$\min_{V_v} \mathcal{J}(\mathbf{V}_t) = \left\|\mathbf{X}_t - \mathbf{U}_t\mathbf{V}_t^T\right\|_F^2 + \delta\left\|\mathbf{V}_t - \mathbf{V}_{t0}\right\|_F^2 \tag{5}$$

where $\delta = \frac{\gamma}{\alpha}$. Taking the derivation of $\mathcal{J}(\mathbf{V}_t)$ and setting it to zero, we can obtain the following form:

$$(-\mathbf{X}_t^T\mathbf{U}_t + \mathbf{V}_t\mathbf{U}_t^T\mathbf{U}_t) + \delta\mathbf{D}_t(\mathbf{V}_t - \mathbf{V}_{t0}) = 0 \tag{6}$$

where $\mathbf{D}_t$ is a diagonal matrix with jth element on the diagonal $\mathbf{D}(j,j) = \frac{1}{2\|\mathbf{V}_t(j,:) - \mathbf{V}_t0(j,:)\|_2}$. In Eq. (6), solving $\mathbf{V}_t$ directly is intractable. Since $\mathbf{D}_t$ and $\mathbf{U}_t^T\mathbf{U}_t$ are symmetric and positive definite, we employ eigen decomposition for them as:

$$\begin{aligned} \mathbf{U}_t^T\mathbf{U}_t &= \mathbf{U}_1\mathbf{\Lambda}_1\mathbf{U}_1^T \\ \mathbf{D}_t &= \mathbf{U}_2\mathbf{\Lambda}_2\mathbf{U}_2^T \end{aligned} \tag{7}$$

where $\mathbf{U}_1, \mathbf{U}_2$ are eigen vectors and $\mathbf{\Lambda}_1, \mathbf{\Lambda}_2$ are diagonal matrices with eigen values on the diagonal. Substituting $\mathbf{U}_t^T\mathbf{U}_t$ and $\mathbf{D}_t$ in Eq. (6), we have:

$$\mathbf{V}_t\mathbf{U}_1\mathbf{\Lambda}_1\mathbf{U}_1^T + \delta\mathbf{U}_2\mathbf{\Lambda}_2\mathbf{U}_2^T\mathbf{V}_t = X_t^T\mathbf{U}_t + \delta\mathbf{D}_t\mathbf{V}_{t0} \tag{8}$$

Multiplying $\mathbf{U}_2^T$ and $\mathbf{U}_1$ from left to right on both sides:

$$\mathbf{U}_2^T\mathbf{V}_t\mathbf{U}_1\Lambda_1 + \delta\Lambda_2\mathbf{U}_2^T\mathbf{V}_t\mathbf{U}_1 = \mathbf{U}_2^T(\mathbf{X}_t\mathbf{U}_t + \delta\mathbf{D}_t\mathbf{V}_{t0})\mathbf{U}_1 \tag{9}$$

Let $\widetilde{\mathbf{V}_t} = \mathbf{U}_2^T\mathbf{V}_t\mathbf{U}_1$ and $\mathbf{Q} = \mathbf{U}_2^T(\mathbf{X}_t\mathbf{U}_t + \delta\mathbf{D}_t\mathbf{V}_{t0})\mathbf{U}_1$, Eq. (9) becomes $\widetilde{\mathbf{V}_t}\mathbf{\Lambda}_1 + \delta\mathbf{\Lambda}_2\widetilde{\mathbf{V}_t} = \mathbf{Q}$, then we can obtain the $\widetilde{\mathbf{V}_t}$ and $\mathbf{V}_t$ as:

$$\begin{aligned} \widetilde{\mathbf{V}_t}(s,l) &= \frac{\mathbf{Q}(s,l)}{\delta\lambda_2^s + \lambda_1^l} \\ \mathbf{V}_t &= \mathbf{U}_2\widetilde{\mathbf{V}_t}\mathbf{U}_1^T \end{aligned} \tag{10}$$

where λ_2^s is the s-th eigen value of $\mathbf{D}_t$ and λ_1^l is l-th eigen value of $\mathbf{U}_t^T\mathbf{U}_t$. The following theorem shows that the updating rule in Eq(10) can monotonically decrease the objective function $\mathcal{J}(\mathbf{V}_t)$.

Theorem 1. *The update rule in Eq. (10) can monotonically decrease the value of $\mathcal{J}(\mathbf{V}_t)$*

Proof. The proof is similar to that in [Nie *et al.*, 2010], due to space limit, we omit the details of the proof.

Update $\mathbf{V}_v$. If $\mathbf{U}_0$, $\mathbf{U}_t$, $\mathbf{V}_t$ and $\mathbf{U}_v$ are fixed, by setting the derivation of the objective function to zero, $\mathbf{V}_v$ can be easily obtained as $\mathbf{V_v} = \mathbf{X}_v^T\mathbf{U}_v(\mathbf{U}_v^T\mathbf{U}_v)^{-1}$. Moreover, we can easily verify updating $\mathbf{V}_v$ will monotonically decrease the objective function.

Update $\mathbf{U}_v$: If $\mathbf{V}_v$, $\mathbf{U}_t$, $\mathbf{V}_t$ and $\mathbf{U}_0$ are fixed, $\mathbf{U}_v$ can be obtained by the following optimization problem:

$$\begin{aligned} \min_{\mathbf{U}_v} & \mathcal{J}(\mathbf{U}_v) = \left\|\mathbf{X}_v - \mathbf{U}_v\mathbf{V}_v^T\right\|_F^2 + \beta\left\|\mathbf{U}_v - \mathbf{U}_0\right\|_F^2 \\ & s.t. \qquad \mathbf{U}_v \geq 0 \end{aligned} \tag{11}$$

The Lagrangian function of Eq. (11) is :

$$\min_{\mathbf{U}_v}\mathcal{L}(\mathbf{U}_v) = \left\|\mathbf{X}_v - \mathbf{U}_v\mathbf{V}_v^T\right\|_F^2 + \beta\left\|\mathbf{U}_v - \mathbf{U}_0\right\|_F^2 - Tr(\Gamma\mathbf{U}_v) \tag{12}$$

where Γ is Lagrangian multiplier. Taking the deviation of $\mathcal{J}(\mathbf{U}_v)$ and using the KKT condition ($\Gamma(s,l)U_v(s,l) = 0$), we can obtain:

$$(-\mathbf{X}_v\mathbf{V}_v + \mathbf{U}_v\mathbf{V}_v^T\mathbf{V}_v + \beta\mathbf{U}_v - \beta\mathbf{U}_0)_{sl}(\mathbf{U}_v)_{sl} = 0 \tag{13}$$

which leads to the following update rule for $\mathbf{U}_v$:

$$(\mathbf{U}_v)_{sl} \leftarrow (\mathbf{U}_v)_{sl}\sqrt{\frac{((\mathbf{X}_v\mathbf{V}_v)^+ + \mathbf{U}_v(\mathbf{V}_v^T\mathbf{V}_v)^- + \beta\mathbf{U}_0)_{sl}}{((\mathbf{X}_v\mathbf{V}_v)^- + \mathbf{U}_v(\mathbf{V}_v^T\mathbf{V}_v)^+ + \beta\mathbf{U}_v)_{sl}}} \tag{14}$$

where $(\mathbf{X}(s,l))^+ = (|\mathbf{X}(s,l)| + \mathbf{X}(s,l))/2$, $(\mathbf{X}(s,l))^- = (|\mathbf{X}(s,l)| - \mathbf{X}(s,l))/2$ and $\mathbf{X} = \mathbf{X}^+ - \mathbf{X}^-$.

Theorem 2. *Let*

$$\begin{aligned}
H(\mathbf{U}_v) &= Tr(-2\mathbf{X}_v\mathbf{V}_v\mathbf{U}_v^T + \mathbf{U}_v\mathbf{V}_v^T\mathbf{V}_v\mathbf{U}_v^T) \\
&\quad + \beta Tr(-2\mathbf{U}_v^T\mathbf{U}_0 + \mathbf{U}_v^T\mathbf{U}_v) \\
h(\mathbf{U}_v, \widetilde{\mathbf{U}_v}) &= \sum_{sl}((\mathbf{X}_v\mathbf{V}_v)^-(s,l)\frac{\widetilde{\mathbf{U}_v^2}(s,l) + \mathbf{U}_v^2(s,l)}{\widetilde{\mathbf{U}_v}(s,l)} \\
&\quad + \beta\frac{\widetilde{\mathbf{U}_v}(s,l)\mathbf{U}_v^2(s,l)}{\widetilde{\mathbf{U}_v}(s,l)} \\
&\quad + (\mathbf{V}_v^T\mathbf{V}_v)^+(s,l)\frac{\widetilde{\mathbf{U}_v}(s,l)\mathbf{U}_v^2(s,l)}{\widetilde{\mathbf{U}_v}(s,l)}) \\
&\quad - \sum_{sl}(2(\mathbf{X}_v\mathbf{V}_v)^+\widetilde{\mathbf{U}_v}(s,l)(1 + \log\frac{\mathbf{U}_v(s,l)}{\widetilde{\mathbf{U}_v}(s,l)}) \\
&\quad + 2\beta\mathbf{U}_0(s,l)\widetilde{\mathbf{U}_v}(s,l)(1 + \log\frac{\mathbf{U}_v(s,l)}{\widetilde{\mathbf{U}_v}(s,l)})) \\
&\quad - \sum_{k,s,l}(\mathbf{V}_v^T\mathbf{V}_v)^-(s,l)\widetilde{\mathbf{U}_v}(k,s)\widetilde{\mathbf{U}_v}(k,l) \\
&\quad (1 + \log\frac{\mathbf{U}_v(k,s)\mathbf{U}_v(k,l)}{\widetilde{\mathbf{U}_v}(k,s)\widetilde{\mathbf{U}_v}(k,l)})
\end{aligned} \tag{15}$$

The auxiliary function $h(\mathbf{U}_v\widetilde{\mathbf{U}_v})$ of $H(\mathbf{U}_v)$ is convex and the global minimum of $h(\mathbf{U}_v, \widetilde{\mathbf{U}_v})$is:

$(\mathbf{U}_v)_{sl} \leftarrow (\mathbf{U}_v)_{sl}\sqrt{\frac{((\mathbf{X}_v\mathbf{V}_v)^+ + \mathbf{U}_v(\mathbf{V}_v^T\mathbf{V}_v)^- + \beta\mathbf{U}_0)_{sl}}{((\mathbf{X}_v\mathbf{V}_v)^- + \mathbf{U}_v(\mathbf{V}_v^T\mathbf{V}_v)^+ + \beta\mathbf{U}_v)_{sl}}}$

Proof: The proof is similar to [Ding *et al.*, 2006] and [Ding *et al.*, 2010], due to space limit, we omit the details.

Theorem 3. *Updating $\mathbf{U}_v$ in Eq. (14) will monotonically decrease the value of objective function $\mathcal{J}(\mathbf{U}_v)$*

Proof: $H(\mathbf{U}_v)$ is the KKT condition of the Lagrangian function for Eq. (11). Based on the definition of auxiliary function and **Theorem 2** we can obtain the following equations:

$$H(\mathbf{U}_v^0) = h(\mathbf{U}_v^0, \mathbf{U}_v^0) \geq h(\mathbf{U}_v^0, \mathbf{U}_v^1) \geq h(\mathbf{U}_v^1, \mathbf{U}_v^1) \geq H(\mathbf{U}_v^1)... \tag{16}$$

This shows the update rule will monotonically decrease the objective function $H(\mathbf{U}_v)$, which complete the proof.

Update $\mathbf{U}_t$: It is worth noting that the procedure of solving $\mathbf{U}_t$ is exactly the same as that of $\mathbf{U}_v$. Thus, we omit the solution of $\mathbf{U}_t$ here.

Update $\mathbf{U}_0$: With $\mathbf{U}_v$, $\mathbf{U}_t$, $\mathbf{V}_t$ and $\mathbf{V}_v$ fixed, the sentiment label $\mathbf{U}_0$ can be obtained by solving the following optimization problem:

$$\min_U \mathcal{J}(\mathbf{U}_0) = \|\mathbf{U}_v - \mathbf{U}_0\|_F^2 + \|\mathbf{U}_t - \mathbf{U}_0\|_F^2 \quad s.t. \quad \mathbf{U}_0^T\mathbf{U}_0 = I; \mathbf{U}_0 \geq 0 \tag{17}$$

The Lagrangian function of Eq. (17) is:

$$\min_U \mathcal{J}(\mathbf{U}_0) = \|\mathbf{U}_v - \mathbf{U}_0\|_F^2 + \|\mathbf{U}_t - \mathbf{U}_0\|_F^2 + Tr(\Lambda(\mathbf{U}_0^T\mathbf{U}_0 - I)) - Tr(\Gamma\mathbf{U}_0) \tag{18}$$

where Λ and Γ are Lagrangian multipliers. Taking the derivation of $\mathcal{J}(\mathbf{U}_0)$ and using KKT conditions we can obtain

$$(\mathbf{U}_0 - \mathbf{U}_v + \mathbf{U}_0 - \mathbf{U}_t + \mathbf{U}_0\Lambda)_{sl}(U_0)_{sl} = 0 \tag{19}$$

which leads the following update rule for $\mathbf{U}_0$:

$$(\mathbf{U}_0)_{sl} \leftarrow (\mathbf{U}_0)_{sl}\sqrt{\frac{(\mathbf{U}_v + \mathbf{U}_t + (\mathbf{U}_0\mathbf{\Lambda})^-)_{sl}}{((\mathbf{U}_0\mathbf{\Lambda})^+ + 2\mathbf{U}_0)_{sl}}} \tag{20}$$

Note that updating $\mathbf{U}_0$ needs updating the Lagrangian multiplier Λ as well. To obtain Λ, we sum over s and get $\Lambda(s,s) = (\mathbf{U}_0^T\mathbf{U}_v - I + \mathbf{U}_0^T\mathbf{U}_t - I)_{s,s}$. The offdiagonal elements of Λ are approximately obtained from non-negative value of U_0, leading to $\Lambda(s,t) = (\mathbf{U}_0^T\mathbf{U}_v - I + \mathbf{U}_0^T\mathbf{U}_t - I)_{st}$. Overall, we can obtain Λ by combining the diagonal values and off-diagonal values.

With the update rules for all the components in the proposed model, we summarize the solution in Algorithm 1. The convergence of Algorithm 1 is demonstrated as below:

Theorem 4.*With Algorithm 1, the objective function Eq. (4) will converge.*

Proof From **Theorem 1** and **Theorem 2**, the object function monotonically decreases:

$$\mathcal{J}(\mathbf{V}_v^0, \mathbf{U}_v^0) \geq \mathcal{J}(\mathbf{V}_v^1, \mathbf{U}_v^0) \geq \mathcal{J}(\mathbf{V}_v^1, \mathbf{U}_v^1)\mathcal{J}(\mathbf{V}_v^2, \mathbf{U}_v^1)... \geq 0 \tag{21}$$

Similarly, we can have the inequality chain for $\mathcal{J}(V_t, U_t)$. Thus we complete the proof.

4 Experiments

In this section, we conduct experiments to answer the following questions - (1) can the proposed framework do sentiment analysis in an unsupervised scenario? and (2) how does the textual information affect the performance of the proposed framework? We begin by giving details about the experimental settings.

4.1 Experiment Settings

We collect datasets from Flickr and Instagram for this study and we give more details below,

Algorithm 1 The proposed USEA

Input: $\{\mathbf{X}_v, \mathbf{X}_t, \mathbf{V}_{t0}\}$ α, β, γ
Output: k sentiment label for each data instance.
Initialization: $\mathbf{U_t}, \mathbf{U_v}, \mathbf{V_v}, \mathbf{V_t}$
while Not Converge **do**
Update $\mathbf{V}_t$ using Eq.(10) and compute $\mathbf{V}_v = \mathbf{X}_v^T\mathbf{U}_v(\mathbf{U}_v^T\mathbf{U}_v)^{-1}$.
Computing $(\mathbf{X}_v\mathbf{V}_v)^{+,-}$, $(\mathbf{X}_t\mathbf{V}_t)^{+,-}$, $(\mathbf{V}_v^T\mathbf{V}_v)^{+,-}$ and $(\mathbf{V}_t^T\mathbf{V}_t)^{+,-}$
Update $\mathbf{U}_v$ using Eq. (14), similarly update $\mathbf{U}_t$
Computing Λ
Update $\mathbf{U}_0$
End
Using max-pooling for $\mathbf{U}_0$ to predict sentiment labels.

Flickr: On Flickr, an image-hosting Website, users can provide tags and descriptions for each uploaded image. Thus the textual information could be comments, image caption, user profile and tags. The collection of Flickr dataset is based on the image id provided by [Yang *et al.*, 2014], which contains 350,4192 images from 4807 users. Some images are unavailable when we crawled the data; hence we limit the number of images from one user as 50, which leads to a dataset with 140,221 images from 4341 users.

Instagram: Instagram is a service supporting photo-sharing via mobile app, where users take pictures and share them on social networking platforms like Facebook and Twitter. Similar to Flickr, we crawl at most 50 images for each user and get totally 131,224 images from 4853 users. Although the textual information as same as that on Flickr, for some images the number of comments is much bigger than that in Flickr, e.g., the images from celebrities usually contain thousands of comments, and we only consider the latest 50 comments for each image in Instagram.

Establishing Ground Truth: For evaluation purpose, we need to create sentiment labels of images. We follow the scheme in [Yang *et al.*, 2014; Liu, 2012] and create labels for images via images' tags. Since we use tags to create labels of images, we do not consider tag information as textual information in the proposed framework. Labeling each post solely relying on tags may cause noise in the ground truth. Therefore we additionally select 20000 images from Flickr and ask three human subjects to manually create labels for them.

Feature extraction: the proposed method has the ability to incorporate visual and textual information. For visual information, we follow the recent approaches [Yuan *et al.*, 2013; Borth *et al.*, 2013] by using mid-level visual features. The visual features are extracted by a large-scale visual attribute detectors [Borth *et al.*, 2013] and the feature dimension is 1200. Text-based features are formed by the term frequency in user profiles, image captions and comments. It is worth noting that textual features, which contain user descriptions, friends' comments and image captions, are preprocessed by stop word removing and stemming. MPQA[5] lexicon is employed as sentiment signals.

[5] http://mpqa.cs.pitt.edu/

The proposed framework USEA is compared with the following sentiment analysis algorithms:

- **Senti API**:[6]. This API is natural language processing API that performs unsupervised sentiment prediction using word-based sentiment. The method only uses textual information.
- **Sentibank**: As a mid-level visual feature based sentiment analysis approach, it uses large-scale visual attribute detectors and low-level visual features to form the Adjective and Nouns visual sentiment description pairs [Borth *et al.*, 2013].
- **EL**: A topical graphical model based sentiment analysis approach, which models the sentiment by low-level visual features and friends information [Yang *et al.*, 2014].
- **USEA-T**: A variant of the proposed method that only considers the textual information including user profiles, image captions and friends' comments.
- **Random**: It predicts sentiment labels of images by randomly guessing.

Noting that SentiBank [Borth *et al.*, 2013] and EL[Yang *et al.*, 2014] are originally proposed for supervised sentiment analysis. We extend them to unsupervised scenarios by replacing original classifiers such as SVM or logistic regression with K-means. However, the clusters identified by K-means have no sentiment labels and we determine their sentiment labels with the Euclidean distance to the ground truth. We use SentiBank-K, and EL-K to represent these modifications.

4.2 Performance Evaluation

Table 1 lists the comparison results and we make several key observations:

- Most of the time, textual based approaches obtain slight better performance than Random. These results support - (1) textual information is often incomplete and noisy and thus often inadequate to support independent sentiment analysis; and (2) textual information contains important cues for sentiment analysis.
- The proposed framework often obtains better performance than baseline methods. There are two major reasons. First textual information provides semantic meanings and sentiment signals for images. Second we combine visual and textual information for sentiment analysis. The impact of textual information on the proposed framework will be discussed in the following subsection.

In summary, compared to the performance Random, the proposed framework can significantly improve the sentiment analysis performance in a unsupervised scenario.

4.3 Impact of Textual Information

We introduce two parameters α and γ to control contributions from textual information. In this subsection, we investigate the impact of textual information on the proposed framework by examining how the performance of USEA varies with the changes of these parameters.

[6] http://sentistrength.wlv.ac.uk/

Table 1: The comparison results of different methods for sentiment analysis.

Method	Flickr (#20,000)	Flickr (#140,221)	Instagram (#131,224)
Senti API	32.30%	34.15%	37.80%
SentiBank-K	41.32%	41.12%	46.31%
EL-K	36.39%	42.90%	43.21%
USEA-T	37.90%	40.22%	36.41%
USEA	**55.22%**	**56.18%**	**59.94%**
Random	32.81%	33.12%	33.05%

To study the impact of α, we fix $\gamma = 0.7$ and vary the value of α as $\{0.001, 0.1, 0.2, 0.3, 0.5, 0.7, 1.5, 2, 10\}$. The performance variance of USEA w.r.t. α is demonstrated in Figure 2. Note that we only show results in Flickr with manual labels since we have similar observations for other datasets. In general, with the increase of α, the performance first increases greatly, reach its peak value and then decrease dramatically. When we increase α from 0.001 to 0.1, the performance increases from 43.21% to 48.07%, which suggests the importance of textual information. With larger values of α (> 1.5), textual information dominates the learning process and the learnt parameters may overfit.

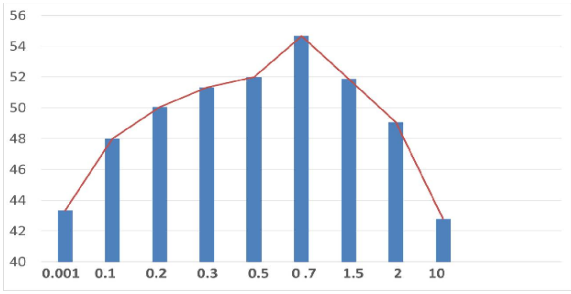


Figure 2: Performance variance w.r.t. α. Y axis is the accuracy performance and X axis is the value of α.

Similarly, to study the impact of γ, we fix $\alpha = 0.7$ and vary the value of γ as $\{0.1, 0.2, 0.3, ..., 0.9, 1\}$. The performance variance of USEA w.r.t. γ is demonstrated in Figure 3. We also only show results in Flickr with manual labels since similar observations are made for other datasets. When γ increases from 0.1 to 0.6, the performance increases a lot, which further supports the importance of sentiment signals from textual information. After 0.8, the increase of γ will reduce the performance dramatically because the proposed framework may overfit to sentiment signals from textual information.

5 Related Work

Recently sentiment analysis have shown success in many aspects, e.g., social response to special events [Hu *et al.*, 2013b; Fukuhara *et al.*, 2007; Diakopoulos and Shamma, 2010], product reviews [Pang and Lee, 2008; Cui *et al.*, 2006], and opinion mining [Liu, 2012; Hu *et al.*, 2013a; Pang *et al.*, 2002; Pak and Paroubek, 2010; Godbole *et al.*, 2007]. Besides, there have been increasing interests in social media images [Borth *et al.*, 2013; Yang *et al.*, 2014; Jia *et al.*, 2012; Yuan *et al.*, 2013], such as images from Twitter and Flickr. Social media are heterogeneous, containing visual and other types of information. Some of existing methods use mainly textual information. For example, [Hu *et al.*, 2013b] proposes a method by counting the word frequency in the user description and predict the sentiment by measuring the word's sentiment. In [Yang *et al.*, 2014], it was argued that friends' comments are more related to the user's sentiment. There are also methods that use solely visual information. For example, [Borth *et al.*, 2013; Yuan *et al.*, 2013; Chen *et al.*, 2014] employ mid-level attributes to model visual content, [Jia *et al.*, 2012] provides a method based on low-level visual features, and [Wang *et al.*, 2015] uses a regulated matrix factorization approach. Inspired by the success of deep learning, [You *et al.*, 2015; Xu *et al.*, 2014] employ a convolution neural network architecture for visual sentiment analysis. However, as discussed previously, these approaches are largely supervised, which means their performance is linked to the assumed availability of a good training set with labels.

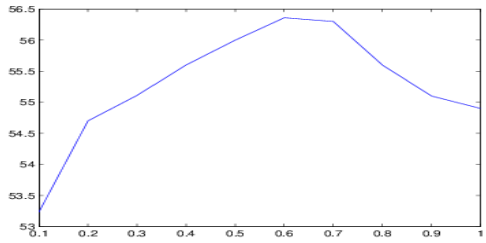


Figure 3: Performance Variance w.r.t. γ. Y axis is the accuracy performance and X axis is the value of γ.

6 Conclusion

In this paper, we propose a novel unsupervised sentiment analysis framework USEA by leveraging textual information and visual information in a unified model. Moreover, USEA provides a new viewpoint for us to better understand how textual information helps bridge the "semantic gap" between visual feature and image sentiment. Experiments on three large-scale datasets demonstrated 1) the advantages of the proposed methods in unsupervised sentiment analysis; and 2) the importance of textual information. In the future, we will exploit more social media sources, such as link information, user history, geo-location, etc., for sentiment analysis.

7 Acknowledgments

Yilin Wang and Baoxin Li are supported in part by National Science Foundation (NSF) under grant number #1135616. Suhang Wang and Huan Liu are supported by, or in part by, the National Science Foundation (NSF) under grant number #1217466 and the U.S. Army Research Office (ARO) under contract/grant number #025071. Any opinions expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies

References

[Borth *et al.*, 2013] Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 223–232. ACM, 2013.

[Chen *et al.*, 2014] Tao Chen, Felix X Yu, Jiawei Chen, Yin Cui, Yan-Ying Chen, and Shih-Fu Chang. Object-based visual sentiment concept analysis and application. In *Proceedings of the ACM International Conference on Multimedia*, pages 367–376. ACM, 2014.

[Cui *et al.*, 2006] Hang Cui, Vibhu Mittal, and Mayur Datar. Comparative experiments on sentiment classification for online product reviews. In *AAAI*, volume 6, pages 1265–1270, 2006.

[Diakopoulos and Shamma, 2010] Nicholas A Diakopoulos and David A Shamma. Characterizing debate performance via aggregated twitter sentiment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1195–1198. ACM, 2010.

[Ding *et al.*, 2006] Chris Ding, Tao Li, Wei Peng, and Haesun Park. Orthogonal nonnegative matrix t-factorizations for clustering. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 126–135. ACM, 2006.

[Ding *et al.*, 2010] Chris Ding, Tao Li, and Michael I Jordan. Convex and semi-nonnegative matrix factorizations. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(1):45–55, 2010.

[Fukuhara *et al.*, 2007] Tomohiro Fukuhara, Hiroshi Nakagawa, and Toyoaki Nishida. Understanding sentiment of people from news articles: Temporal sentiment analysis of social events. In *ICWSM*, 2007.

[Godbole *et al.*, 2007] Namrata Godbole, Manja Srinivasaiah, and Steven Skiena. Large-scale sentiment analysis for news and blogs. *ICWSM*, 7:21, 2007.

[Hu and Liu, 2004] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM, 2004.

[Hu *et al.*, 2013a] Xia Hu, Jiliang Tang, Huiji Gao, and Huan Liu. Unsupervised sentiment analysis with emotional signals. In *Proceedings of the 22nd international conference on World Wide Web*, pages 607–618. International World Wide Web Conferences Steering Committee, 2013.

[Hu *et al.*, 2013b] Yuheng Hu, Fei Wang, and Subbarao Kambhampati. Listening to the crowd: automated analysis of events via aggregated twitter sentiment. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pages 2640–2646. AAAI Press, 2013.

[Jia *et al.*, 2012] Jia Jia, Sen Wu, Xiaohui Wang, Peiyun Hu, Lianhong Cai, and Jie Tang. Can we understand van gogh's mood?: learning to infer affects from images in social networks. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 857–860. ACM, 2012.

[Liu, 2012] Bing Liu. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167, 2012.

[Nie *et al.*, 2010] Feiping Nie, Heng Huang, Xiao Cai, and Chris H Ding. Efficient and robust feature selection via joint 2, 1-norms minimization. In *Advances in Neural Information Processing Systems*, pages 1813–1821, 2010.

[Pak and Paroubek, 2010] Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *LREC*, volume 10, pages 1320–1326, 2010.

[Pang and Lee, 2008] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135, 2008.

[Pang *et al.*, 2002] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics, 2002.

[Wang *et al.*, 2015] Yilin Wang, Yuheng Hu, Subbarao Kambhampati, and Baoxin. Li. Inferring sentiment from web images with joint inference on visual and social cues: A regulated matrix factorization approach. In *ICWSM*, page 21, 2015.

[Xu *et al.*, 2014] Can Xu, Suleyman Cetintas, Kuang-Chih Lee, and Li-Jia Li. Visual sentiment prediction with deep convolutional neural networks. *arXiv preprint arXiv:1411.5731*, 2014.

[Yang *et al.*, 2014] Yang Yang, Jia Jia, Shumei Zhang, Boya Wu, Juanzi Li, and Jie Tang. How do your friends on social media disclose your emotions? 2014.

[You *et al.*, 2015] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. Robust image sentiment analysis using progressively trained and domain transferred deep networks. 2015.

[Yuan *et al.*, 2013] Jianbo Yuan, Sean Mcdonough, Quanzeng You, and Jiebo Luo. Sentribute: image sentiment analysis from a mid-level perspective. In *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*, page 10. ACM, 2013.